Swami Ramanand Teerth Marathwada University

Vishnupuri, Nanded - 431 606

Synopsis on

DEVNAGARI TEXT RECOGNITION

Submitted by
Ms. Vina M. Lomte

Under Guidance
of
Dr. D. D. Doye

Department of Computer Science & Engineering
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded

2015-2016

Research Topic	: **DEVNAGARI TEXT RECONITION**
Research Centre	**:** Shri Guru Gobind Singhji Institute of Engineering and Technology, Vishnupuri, Nanded (Maharashtra State) Pin 431 606 INDIA
University	**: Swami Ramanand Teerth Marathwada University, Vishnupuri, Nanded - 431 606**
Research Scholar	**: Ms. Vina. M. Lomte**
Guide	**: Dr. D. D. Doye**

Ms. Vina. M. Lomte **Dr. D. D. Doye**

PROBLEM STATEMENT

To design a system which will be efficient and accurate to recognize characters, joint characters, numbers and special characters of Devanagari script.

ABSTRACT

Recognition of handwritten language words and characters have always been a challenging field for computing field. In this work we deal with problem of recognition of Devanagari characters, joint characters, numbers and special characters. We choose Devanagari script as our problem script as it is widely used by the population in many states in India. If this problem is solved it would be of much help to archaeologists, historical document recognition and for most of general purpose document recognition. Much work have been done on character recognition of Devnagari script but very little have been done on joint characters, numbers and special characters recognition. Some methods are currently in use for character recognition but they lack efficiency and accuracy giving us the scope to improve it further, also the work for recognition of joint characters, numbers and special characters is still under research. In this work we try to provide a solution to the above mentioned research problems.

INTRODUCTION

Even though a descendent of the Brahmi script, Devanagari has evolved into a highly cursive script. Many languages in India, such as Hindi and Sanskrit, use Devanagari and many more languages throughout India use local variants of this script. Hindu scriptures are written in Devanagari, a fact illustrated by the etymology of the name. "Devanagari" is a compound word with two roots: deva means "deity", and nagari means "city". Together it implies a script that is both religious as well as urbane or sophisticated The Devanagari script is used for over 120 languages including Hindi, Marathi, Nepali, Pali, Konkani, Bodo, Sindhi and Maithili among other languages and dialects, making it one of the most used and adopted writing systems in the world. The Devanagari script is also used for classical Sanskrit texts. The Devanagari script is closely related to the Nandinagari script commonly found in numerous ancient manuscripts of South India, and it is distantly related to a number of Southeast Asian scripts

Devnagari is spoken by more than 500 million people in the world. It is 4th major spoken language in the world after Chinese, Spanish and English.It can be utilized to also recognize ancient texts which are mostly in sanskrit and pali, which in turn can be used to recognize several ancient texts of archealogy which will unfold many mistries of HIndu culture. Till now only handwritten characters of Devnagri script can be recognized by present methods but an approach to recognize words and sentences is required to utilize the technology to the fullest. Devnagri script has many combined sounds made by combining two or more vowels and consonants or two or more consonants together also numericals are yet to be recognized.

OBJECTIVES

a) To improve accuracy of current techniques.
b) To recognize all special characters, numbers and joint characters in Devnagri script.
c) To increase dataset which in turn will result in improved efficiency.
d) To design a system which can be utilized by majority of people using Devnagri script.

LITERATURE SURVEY

1. Toselli *et-al* did Handwritten Text Recognition experiments and results are presented on the historical Bentham text image dataset used in the ICFHR-2014 HTRtS competition. The successful segmentation-free holistic framework is adopted, using traditional modeling approaches based on Hidden Markov optical character models (HMM) and an N-gram language model (LM). Departing from the very basic N-gram-HMM baseline system provided in HTRtS, several improvements are made in text image feature extraction and LM and HMM modeling, including more accurate HMM training by means of discriminative training. As a result, they achieved similar recognition accuracy as some of the best performing (single, uncombined) systems based on (recurrent) Neural Networks (NN), using identical training and testing data. The traditional N-gram/HMM framework offers several advantages over modern approaches based on (hybrid, recurrent) NNs. Perhaps the most important is the much faster training of HMMs and the well understood stability of the results of Baum-Welch training. These advantages become very important when dealing with many historical document collections, which are typically huge and entail very high degrees of variability, making it generally difficult to re-use models trained on previous collections.[1]

2. Sundaram *et-al* in this article describes a post processing strategy for online, handwritten, isolated Tamil words. Contributions have been made with regard to two issues hardly addressed in the online Indic word recognition literature, namely, use of (1) language models exploiting the idiosyncrasies of Indic scripts (2) expert classifiers for the disambiguation of confused symbols.The input word is first segmented into its individual symbols, which are recognized using a primary support vector machine (SVM) classifier. Thereafter, we enhance the recognition accuracy by utilizing (i) a bigram language model at the symbol or character level and (ii) expert classifiers for reevaluating and disambiguating the different sets of confused symbols. The symbol-level bigram model is used in a traditional Viterbi framework. The concept of a character comprising multiple symbols is unique to Dravidian languages such as Tamil. This multi-symbol feature of Tamil characters has been exploited in proposing a novel, prefix-tree-based character-level bigram model that does not use Viterbi search; rather it reduces the search space for each input symbol based on its left context. For disambiguating confused symbols, a dynamic time-warping approach is proposed to automatically identify the parts of the online trace that discriminates between the confused classes. Fine classification of these regions by dedicated expert SVMs reduces the extent of confusions between such symbols. The integration of segmentation, prefix-tree-based language model and disambiguation of confused symbols is presented on a set of 15,000 handwritten isolated online Tamil words. Their results show recognition accuracies of 93.0% and 81.6% at the symbol and word level, respectively, as compared to the baseline classifier performance of 88.4% and 65.1%, respectively.[2]

3. Yacoubi *et-al* in This paper describes a system that recognizes freely handwritten words off-line. Based on Hidden Markov models (HMMs), this system is designed in the context of a real application in which the vocabulary is large but dynamically limited. After preprocessing, a word image is segmented into letters or pseudo-letters and represented by two feature sequences of equal length, each consisting of an alternating sequence of shape-symbols and segmentation-symbols that are both explicitly modeled. The word model is made up of the concatenation of appropriate letter models which

consist of elementary HMMs. Two mechanisms are considered to reject unreliable outputs, depending on whether or not the unknown word image is guaranteed to belong to the dynamic lexicon. Experiments performed on real data show that HMMs can be successfully used for handwriting recognition.[3]

4. Lavrenko *et-al* most offline handwriting recognition approaches proceed by segmenting words into smaller pieces (usually characters) which are recognized separately. The recognition result of a word is then the composition of the individually recognized parts. Inspired by results in cognitive psychology, researchers have begun to focus on holistic word recognition approaches. Here they present a holistic word recognition approach for single-author historical documents, which is motivated by the fact that for severely degraded documents a segmentation of words into characters will produce very poor results. The quality of the original documents does not allow us to recognize them with high accuracy their goal here is to produce transcriptions that will allow successful retrieval of images, which has been shown to be feasible even in such noisy environments. Authors believe that this is the first systematic approach to recognizing words in historical manuscripts with extensive experiments. Their experiments show a recognition accuracy of 65%, which exceeds performance of other systems that operate on non-degraded input images (non historical documents).[4]

5. Impedvo *et-al* this paper presents a new approach for Handwritten Word Recognition based on Hidden Markov Model theory and the sliding window technique. The new approach uses specific singularity markers to support the recognition phase: the Static Marker and the Dynamic Marker. Moreover, different strategies for sliding window step are considered: Regular Step and Progressive Step. Experimental results showing the improvements obtained for basic word lexicon recognition are reported in the paper.[5]

6. In this work these researchers said that the chic of writing numbers is diverse and comes in a variety of form, size and fonts. Identification of Indian languages scripts is challenging problems. In Optical Character Recognition [OCR], machine printed or handwritten characters/numerals are recognized. There are plentiful approaches that deal with problem of detection of numerals/character depending on the sort of feature extracted and different way of extracting them. This paper proposes a recognition scheme for handwritten Hindi (devnagiri) numerals; most admired one in Indian subcontinent their work focused on a technique in feature extraction i.e. Local-based approach, a method using 16-segment display concept, which is extracted from halftoned images & Binary images of isolated numerals. These feature vectors are fed to neural classifier model that has been trained to recognize a Hindi numeral. The archetype of system has been tested on varieties of image of numerals. Experimentation result shows that recognition rate of halftoned images is 98 % compared to binary images (95%).[6]

7. Dixit *et-al* gives a new approach for recognition of handwritten Devanagari characters. Twenty handwritten characters from 100 people resulting 2000 characters are used for the experimentation. The handwritten characters written of paper are scanned, preprocessed and on every individual characters wavelet transform is applied so as to get decomposed images of characters. Statistical parameters are computed over the decomposition to form feature vector. The feature vectors serve as input to back propagation neural networks for classification into one of 20 classes and based classes

they are recognized. The accuracy obtained is around 70 percent over large number of samples.[7]

8. Belhe *et-al* proposed approach performs recognition of online handwritten isolated Hindi words using a combination of HMMs trained on Devanagari symbols and a tree formed by the multiple, possible sequences of recognized symbols. In general, words in Indic languages are composed of a number of aksharas or syllables, which in turn are formed by groups of consonants and vowel modifiers. Segmentation of aksharas is critical to accurate recognition of both recognition primitives as well as the complete word. Also, recognition in itself is an intricate job. This holistic task of akshara segmentation, symbol identification and subsequent word recognition is targeted in their work. It is handled in an integrated segmentation-recognition framework. By making use of online stroke information for postulating symbol candidates and deriving HOG feature set from their image counterparts, the recognition becomes independent of stroke order and stroke shape variations. Thus, the system is well suited to unconstrained handwriting. Data for this work is collected from different parts of India where Hindi language is predominantly in use. Symbols extracted from 60,000 words are used to train and test 140 symbol-HMM models. The system is designed to output one or more candidate words to the user, by tracing multiple tree paths (up to leaf nodes) under the condition that the symbol likelihood (confidence score) at every node is above threshold. Tests performed on 10,000 words yield an accuracy of 89%.[8]

9. In this paper Ramachandrula *et-al* discusses the Hindi offline handwritten word recognizer (HWR) that we are developing. For the purpose of training and testing the offline HWR, we have created a Hindi handwritten word and character database from 100 writers. In their HWR we use two-pass Dynamic Programming algorithm to match the test word against each word in the lexicon by initially segmenting the test word image into probable characters. We extract directional element features (DEF) on each character image segment and statistically model them. Currently we are achieving word recognition accuracies of 91.23 % to 79.94 % on 10 to 30 vocabulary words.[9]

10. In this article, Bhowmik *et-al* propose a new framework for segmentation of Bangla handwritten word images into meaningful individual symbols or pseudo-characters. Existing segmentation algorithms are not usually treated as a classification problem. However, in the present study, the segmentation algorithm is looked upon as a two-class supervised classification problem. The method employs an SVM classifier to select the segmentation points on the word image on the basis of various structural features. For training of the SVM classifier, an unannotated training set is prepared first using candidate segmenting points. The training set is then clustered, and each cluster is labeled manually with minimal manual intervention. A semi-automatic bootstrapping technique is also employed to enlarge the training set from new samples. The overall architecture describes a basic step toward building an annotation system for the segmentation problem, which has not so far been investigated. The experimental results show that their segmentation method is quite efficient in segmenting not only word images but also handwritten texts. As a part of this work, a database of Bangla handwritten word images has also been developed. Considering their data collection method and a statistical analysis of their lexicon set, they claim that the relevant characteristics of an ideal lexicon set are present in their handwritten word image database.[10]

11. Keerthi *et-al* in this paper, present a model for writer-independent online handwriting character recognition for the 49 basic Hindi characters. The proposed system is implemented on mobile device using two different approaches namely Principal Component Analysis (PCA) and Dynamic Time Wrapping (DTW). To find the suitability of these two approaches for handheld devices several experiments were conducted and detailed analysis has been made on the obtained results. The results obtained for PCA approach is quite promising than DTW. On an average, recognition accuracy up to 86% is achieved for the PCA approach and up to 66% is achieved for DTW approach, also the time taken for recognition of unknown character is around 0.8sec for PCA approach, and around 51sec for DTW approach, thus the PCA approach is suitable for real-time applications.[11]

12. Prasad *et-al* in this paper implemented Gradient feature extraction technique, which provides more than 94% recognition accuracy. They have acquired 1000 samples of handwritten Hindi characters by initializing the mouse in graphics mode. The 500 samples have been used for training the network (Train Data) and remaining 500 samples have been used for testing the network (Test Data). The system has been trained using several different forms of handwritings provided by both male and female participants of different age groups. Finally, this rigorous training results an automatic HCR system using MLP network. The error backpropagation algorithm has been used to train the MLP network. A comparative analysis was performed by implementing both global input and Gradient feature input. They have concluded that gradient feature extraction technique provides better recognition accuracy with reduced training time.[12]

REFERENCES

[1] Alejandro H. Toselli and Enrique Vidal. 2015. Handwritten Text Recognition Results on the Bentham Collection with Improved Classical N-Gram-HMM methods. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing* (HIP '15). ACM, New York, NY, USA, 15-22. 2015.

[2] Suresh Sundaram and A. G. Ramakrishnan. 2015. Bigram Language Models and Reevaluation Strategy for Improved Recognition of Online Handwritten Tamil Words. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 14, 2, Article 8 (April 2015), 28 pages.

[3] A. El-Yacoubi, R. Sabourin, M. Gilloux and C.Y. Suen "Offline handwritten word recognition using HMM" IEEE Trans. Intel. Transp. Syst., vol. 12, no. 2, pp. 548–562, Jun. 2011.

[4] Victor Lavrenko, Toni M. Rath, and R. Manmatha. 2004. Holistic Word Recognition for Handwritten Historical Documents. In *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)* (DIAL '04). IEEE Computer Society, Washington, DC, USA, 278-290.

[5] Sebastiano Impedovo, Anna Ferrante, Raffaele Modugno, "HMM Based Handwritten Word Recognition System by Using Singularities", ICDAR, 2009, 2013 12th International Conference on Document Analysis and Recognition, 2013 12th International Conference on Document Analysis and Recognition 2009, pp. 783-787,

[6] Banashree, N. P., et al. "OCR for script identification of Hindi (Devnagari) numerals using error diffusion Halftoning Algorithm with neural classifier."Proceedings of World Academy of Science Engineering and Technology. Vol. 20. 2007.

[7] Dixit, Adawait, Ashwini Navghane, and Yogesh Dandawate. "Handwritten Devanagari character recognition using wavelet based feature extraction and classification scheme." India Conference (INDICON), 2014 Annual IEEE. IEEE, 2014.

[8] Swapnil Belhe, Chetan Paulzagade, Akash Deshmukh, Saumya Jetley, and Kapil Mehrotra. 2012. Hindi handwritten word recognition using HMM and symbol tree. In *Proceeding of the workshop on Document Analysis and Recognition* (DAR '12). ACM, New York, NY, USA, 9-14.

[9] Sitaram Ramachandrula, Shrang Jain, and Hariharan Ravishankar. 2012. Offline handwritten word recognition in Hindi. In *Proceeding of the workshop on Document Analysis and Recognition* (DAR '12). ACM, New York, NY, USA, 49-54.

[10] Tapan Kumar Bhowmik, Swapan Kumar Parui, Utpal Roy, and Lambert Schomaker. 2016. Bangla Handwritten Character Segmentation Using Structural Features: A Supervised and Bootstrapping Approach. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 15, 4, Article 29 (April 2016), 26 pages.

[11] Dayashankar Singh, Maitreyee Dutta, and Sarvpal H. Singh. 2009. Neural network based handwritten hindi character recognition system. In *Proceedings of the 2nd Bangalore Annual Compute Conference* (COMPUTE '09). ACM, New York, NY, USA, , Article 15 , 4 pages.

[12] G. Keerthi Prasad, Imran Khan, and Naveen Chanukotimath. 2012. On-line Hindi handwritten character recognition for mobile devices. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (ICACCI '12). ACM, New York, NY, USA, 1074-1078.

Ms. Vina M. Lomte
Researcher

Dr. D. D. Doye
Guide